

## 8.5 Linear Regression





## Goals for the Day

**1**

Scatterplots

**2**

Correlation

**3**

Using Your  
Calculator

**4**

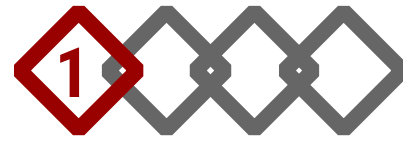
Regression

1

# Scatterplots



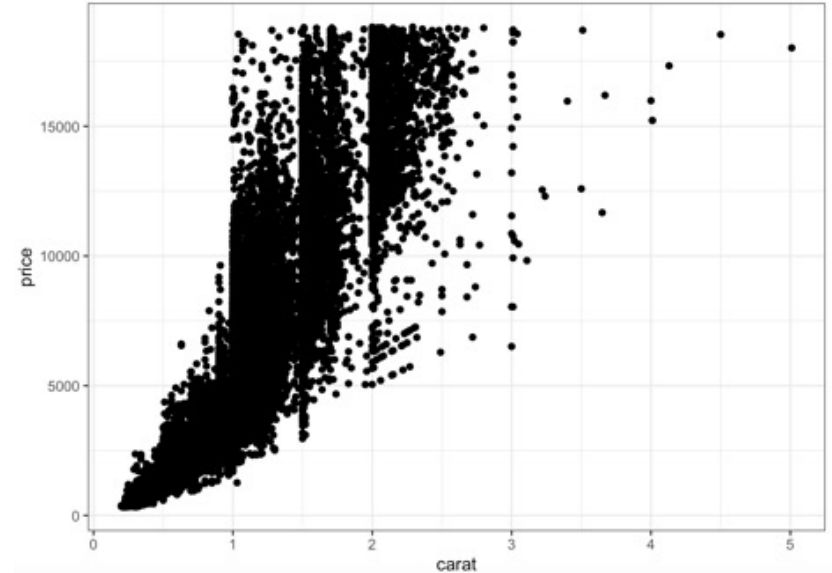
# Scatterplots



Displays the relationship between **two quantitative** variables measured on the same individuals.

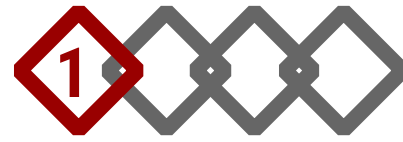
Useful to determine if an **association** exists!

- ▷ So is there a pattern where some values of one variable tend to occur with some values of the other variable.
- ▷ Ex) Smaller carat diamonds tend to have lower prices, and as the carat increases prices tend to increase as well.



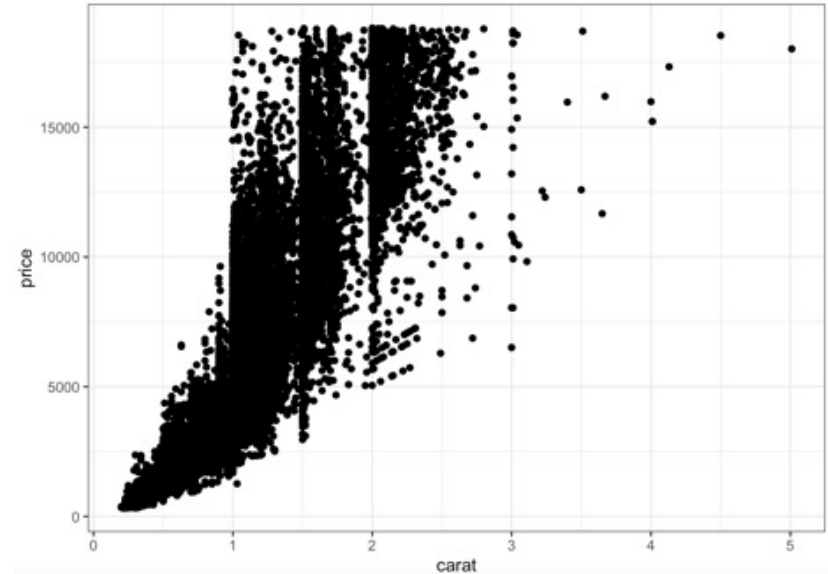


# Scatterplots



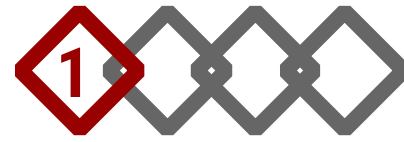
## Setup of axes

- ▷ The explanatory (independent) variable goes on the X (horizontal) axis.
- ▷ The response (dependent) variable goes on the Y (vertical) axis.
- ▷ Ex) How large a diamond is impacts how much it costs → Carat = X; Price = Y.





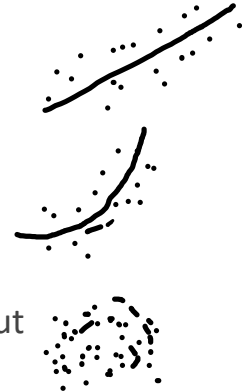
# Interpreting Scatterplots



Interpreting a scatterplot (what we are looking for in a scatterplot)

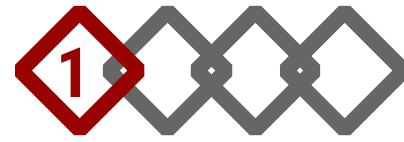
**Form** (pattern of the dots)

- ▷ Linear → Points follow a general linear trend; Straight line.
- ▷ Curved → Points show some evidence of curvature; NOT a straight line.
- ▷ Random scatter → No pattern, points are just scattered about randomly kinda like a cloud of points.





# Interpreting Scatterplots



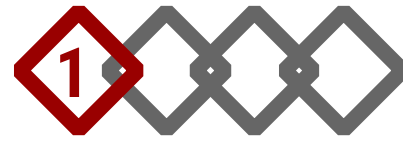
■ Interpreting a scatterplot (what we are looking for in a scatterplot)

■ **Direction** (of the association; only applies to linear relationships)

- ▷ Positive → Upward trend.
- ▷ Negative → Downward trend.
- ▷ No Association → There is no pattern or general trend (corresponds to random scatter).



# Interpreting Scatterplots

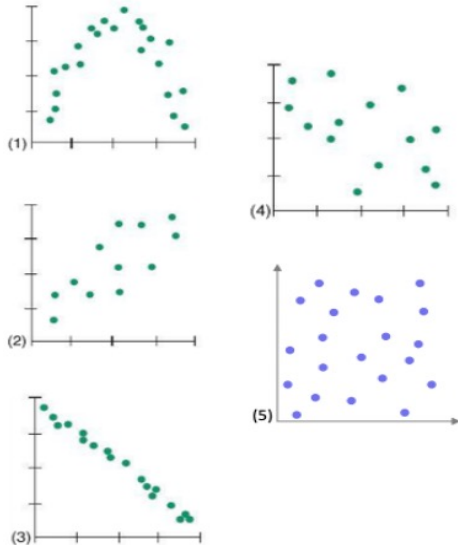
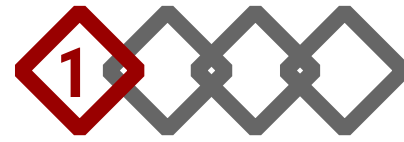


- Interpreting a scatterplot (what we are looking for in a scatterplot)
- **Strength** (how strong the association is; how well the data fits the pattern; only applying this to linear relationships)





# Interpreting Scatterplots Example



## Example

	Form	Direction	Strength
(1)	Curved	N/A (+/-)	N/A (strong)
(2)	Linear	Positive	Moderate
(3)	Linear	Negative	Strong
(4)	Roughly linear	Negative	Weak
(5)	Random scatter	No association	Very weak

# 2

## Correlation



# Correlation

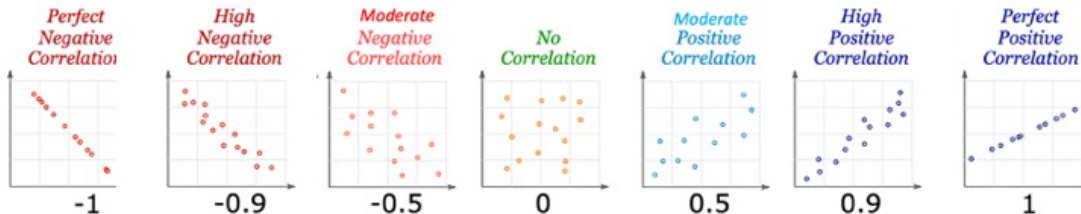
2

■ The **correlation** ( $r$ ) is an index that expresses the direction and strength of the relationship.

- It combines both of these aspects into a single number measure.
- Often referred to as the correlation coefficient (or Pearson's correlation).

■ Interpreting correlation

- Sign = Direction
- Absolute value  $|r|$  = Strength





# Correlation



## Properties of Correlation

- Scale goes from -1 to 1  $\rightarrow -1 \leq r \leq 1$
- Only applies to LINEAR relationships.
- r has no units and is the same regardless of which variable is X or Y.



Does NOT imply a cause-and-effect relationship.

- Ex) Ice cream sales and shark attacks have a strong positive correlation.



# 3

## Using Your Calculator



## Using Your Calculator

Using TI-83/84 (and TI-30 XS MultiView / XIS) to calculate correlation (and regression line).

### Steps for the TI-83/84

1. Enter data: STAT → Edit →  
Enter X data in L<sub>1</sub>  
Enter Y data in L<sub>2</sub>
2. Calculate: STAT → CALC → LinReg(ax+b)
  - a) XList: L<sub>1</sub>.
  - b) YList: L<sub>2</sub>.
  - c) Rest leave blank.
  - d) Calculate!

### Steps for the TI-30XS MultiView

1. Data →  
Enter X data in L<sub>1</sub>  
Enter Y data in L<sub>2</sub>
2. 2<sup>nd</sup> → stat → 2-Var Stats
  - a) xDATA: L1
  - b) yDATA: L2
  - c) CALC

### Steps for the TI-30 XIS

1. 2<sup>nd</sup> → STAT → 2-VAR (Enter)
2. DATA
  - X<sub>1</sub> = # (scroll down)
  - Y<sub>1</sub> = # (scroll down)
  - ... (repeat for all data points)
3. STATVAR (scroll across)
4. To exit this menu: 2<sup>nd</sup> → EXIT STAT → Y

\*\*\* One time setup for TI-83/84: 2<sup>nd</sup> → Catalog → DiagnosticOn → Enter

$$n = 4$$
$$L_1 \quad L_2$$

NORMAL FLOAT AUTO REAL RADIAN HP

**MODE** CALC TESTS

**F1**Edit...

2:SortA(

3:SortD(

4:ClrList

5:SetUpEditor

[illegible]

EDIT **CALC** TESTS  
1:1-Var Stats  
2:2-Var Stats  
3:Med-Med  
4:LinReg(ax+b)  
5:QuadReg  
6:CubicReg  
7:QuartReg  
8:LinReg(a+bx)  
9:LnReg

NORMAL FLOAT AUTO REAL RADIAN HP

**lnRe9(ax+b)**

Xlist:L1  
Ylist:L2  
FreqList:  
Store ResEQ:  
Calculate

NORMAL FLOAT AUTO REAL RADIAN MP

**LinReg**

$y = ax + b$

$a = 1.637931034$  \_\_\_\_\_

$b = 1.103448276$  \_\_\_\_\_

$r^2 = .9634888438$

$r = .9815746756$  \_\_\_\_\_

- Y-intercept

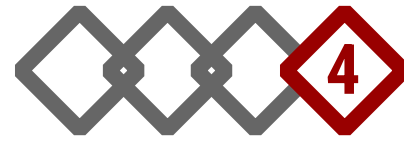
► Correlation

# 4

## Regression

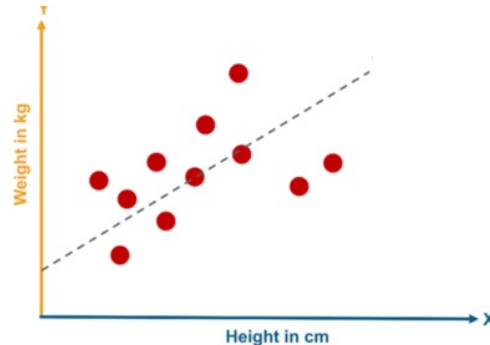


# Regression

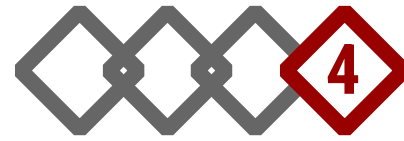


Ultimately, we want to determine if we can use a straight line to model the relationship between two variables → If so, we can use that model to make predictions!

▷ This process is called **Linear Regression**.



# Regression



Step 1 → Determine if there is a **significant correlation** (linear relationship).

a) Calculate correlation.

b) Compare it to the **Table of Critical Values or the Pearson Correlation Coefficient** to see if it is statistically significant.

➤ Match the sample size  $n$  and the Level of significance  $\alpha$  (Probability our claims about the data are wrong) to the specific problem.

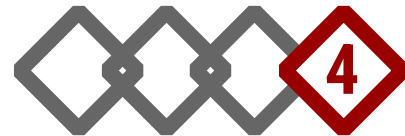
Critical Values of the Pearson Correlation Coefficient		
$n$	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875

★ ➤ If  $|r| > \text{Critical Value (CV)} \rightarrow r$  is statistically significant (unlikely to have occurred by chance).

➤ Demo ex)  $n = 4, \alpha = 0.05 \rightarrow r = 0.982 > 0.950 = \text{CV}$

→ Significant ✓ → Can make predictions ✓

# Regression



■ Step 2 → Once we have a significant correlation, we can find the **regression line**.

- ▷ Linear equation that fits our data best (aka 'line of best fit').
  - ▷ It is IMPORTANT to get the X and Y variables correct!
- ▷ Our calculator gives us our equation!

# Regression

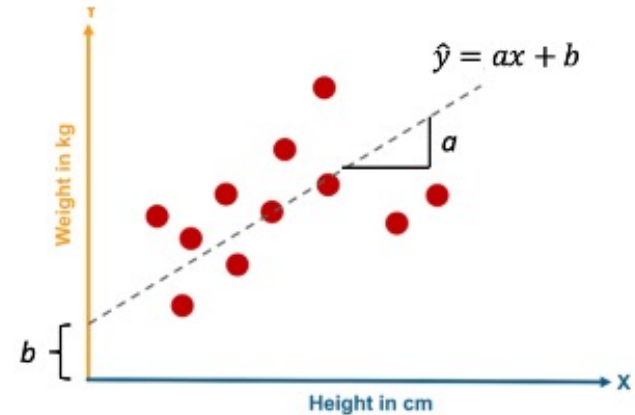
4

## Equation

- Here is the form of our linear equation (written in slope-intercept form):

$$\begin{aligned}\hat{y} &= ax + b && \text{calculator} \\ &= b_0 + b_1x && \text{Hawkes}\end{aligned}$$

- $x$  = Value of the explanatory variable
- $\hat{y}$  = Predicted value of the response variable for the given  $x$
- $a$  (or  $b_1$ ) = Slope
  - It measures the direction and steepness of the line
- $b$  (or  $b_0$ ) = Y intercept
  - It is the location where the regression line crosses the Y-axis (value of Y when X = 0)



# Regression



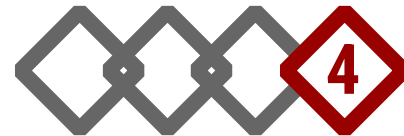
Step 3 → Make **predictions** using the regression line.

- ▶ We can think of our regression line, and specifically  $\hat{y}$ , as predicted values of Y for all X values *in the X range of our sample data!*
- ▶ Calculating these is simple:
  - ▶ Just plug in the new X value to our equation and this will give us the predicted Y.
  - ▶ Demo example) Predict Y for X = 7.

$$\hat{y} = 1.638x + 1.103 \rightarrow \hat{y} = 1.638(7) + 1.103 = 12.569$$

$ax + b$

## Full Example



Hours Spent on Homework	41	20	34	43	9	20	54	52	10	21
Grade on Test	79	63	76	100	55	82	95	80	60	80

- Calculate the correlation for the dataset above and determine if it is statistically significant at a level of significance of  $\alpha = 0.05$ .
- If appropriate, determine the regression equation.
- If a student spends 35 hours on homework, make a prediction for their grade on the test.
- If a student spends 50 hours on homework, make a prediction for their grade on the test.

a)  $r = 0.779 > 0.632 \rightarrow$  Significant

b) Appropriate  $\rightarrow \hat{y} = 0.676x + 50.549$

c)  $\hat{y} = 80.119$

d)  $\hat{y} = 90.259$